StringBio
2018

Statistical Mitogenome Assembly with Repeats

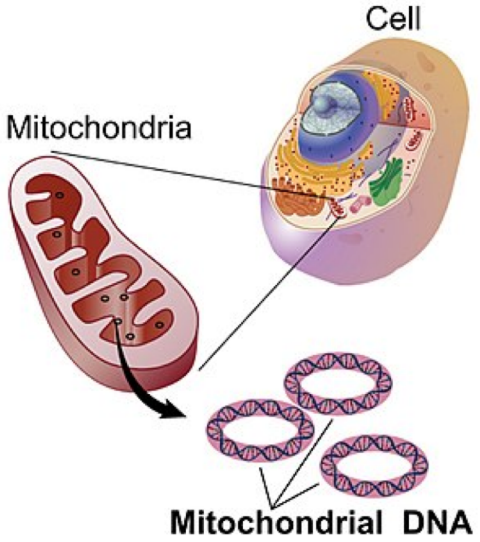
Fahad Alqahtani & Ion Măndoiu

Outline

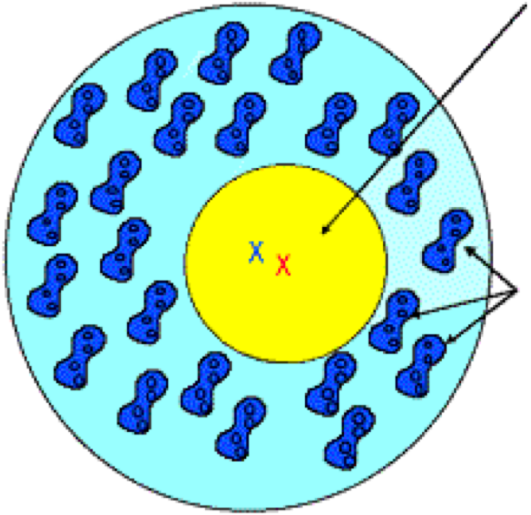
- **Background and prior work**
- SMART pipeline
- Results
- Conclusions and future work

Mitochondria: the powerhouse of the cell

- Cellular organelles within eukaryotic cells
  - Convert chemical energy from food into adenosine triphosphate (ATP)
  - The popular term "powerhouse of the cell" was coined by Philip Siekevitz in 1957



The second genome



Nuclear

- 2 copies/cell
- inherited from both parents
- unique to individual

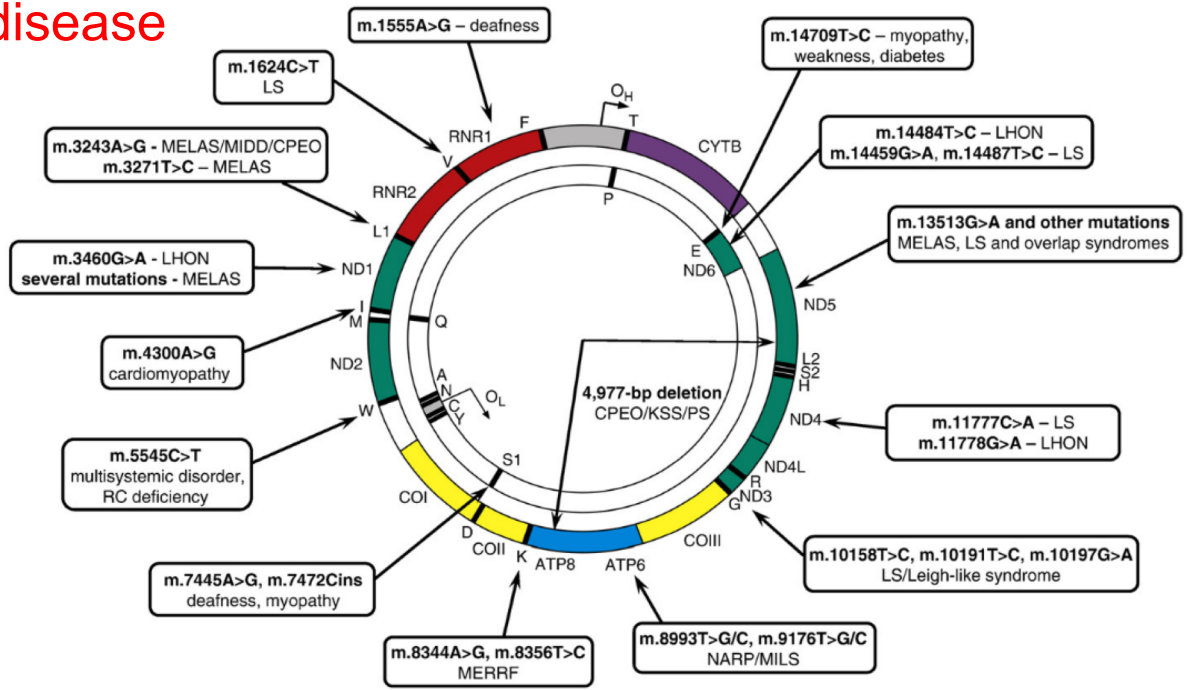
Mitochondrial

- >1000 copies/cell
- maternally inherited
- not unique to individual

Source:https://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july1999/dnalist.htm/dnaf1.htm

Why sequence the mitogenome?

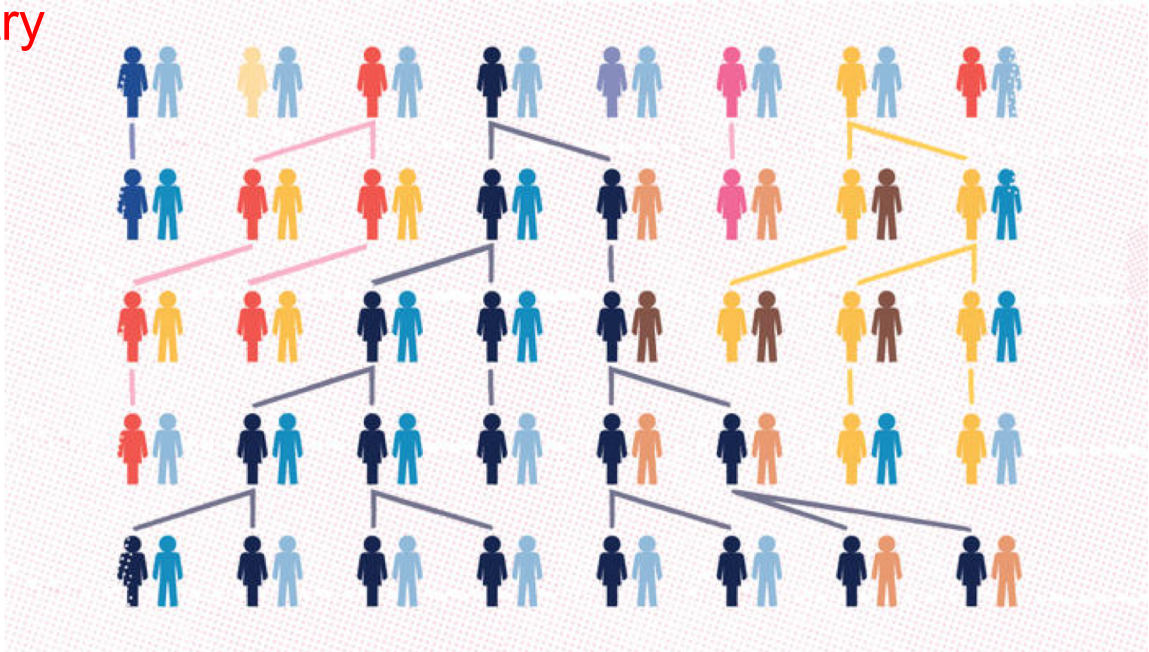
- Important role in disease



Tuppen, Helen AL, et al. "Mitochondrial DNA mutations and human disease." Biochimica et Biophysica Acta (BBA)-Bioenergetics 1797.2 (2010): 113-128.

Why sequence the mitogenome?

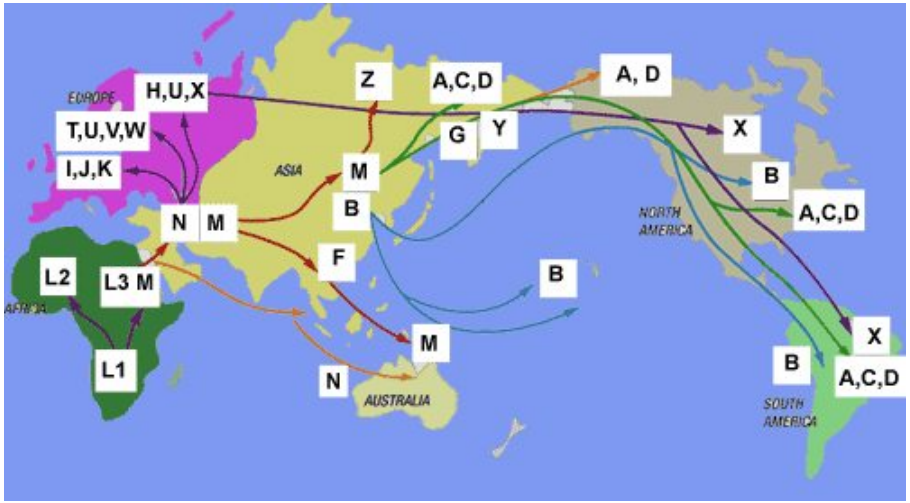
- Important role in disease
- Tracing maternal ancestry



Source: http://www.norwaydna.no/mtdna_en/

Why sequence the mitogenome?

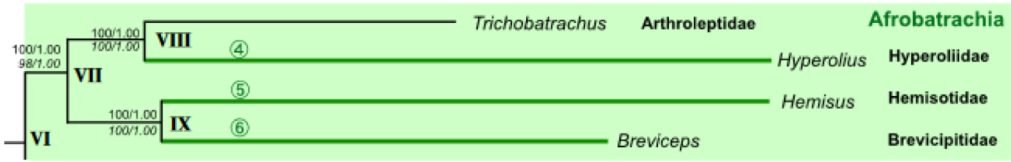
- Important role in disease
- Tracing maternal ancestry
- Inferring human population migrations

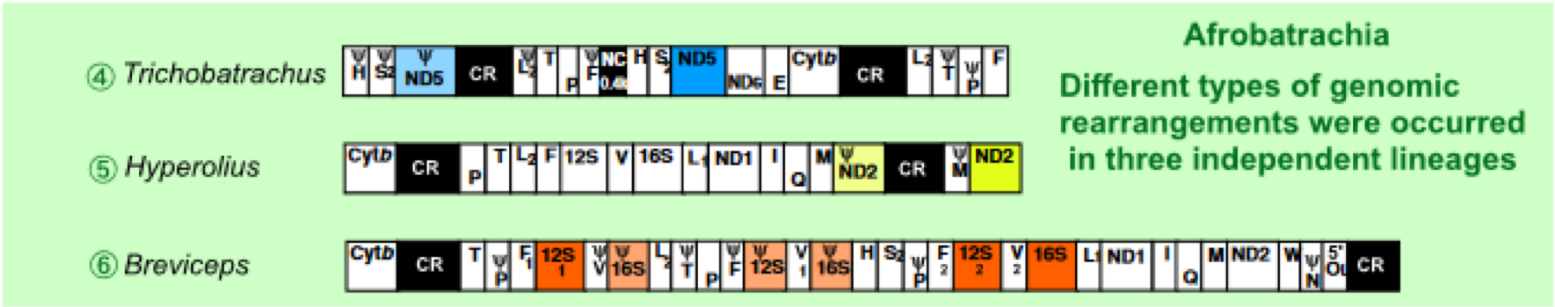


https://blog.23andme.com/ancestry/haplogroups-explained/

Why sequence the mitogenome?

- Important role in disease
- Tracing maternal ancestry
- Inferring human population migrations
- Species tree reconstruction





Kurabayashi, Atsushi, and Masayuki Sumida. "Afrobatrachian mitochondrial genomes: genome reorganization, gene rearrangement mechanisms, and evolutionary trends of duplicated and rearranged genes." BMC genomics 14.1 (2013): 633.

Mitogenome assembly

- Most existing pipelines rely on reference genome or mitogenome of related species
- Off-the-shelf *de novo* assemblers poorly suited for assembling mtDNA from WGS reads
  - Mitochondrial reads often discarded due to much higher sequencing depth of mtDNA compared to gDNA
  - Do not handle well circular genomes & repeats

Prior work

Tool	Method	Input Requirements
MITOBim [Hahn at el 2013]	Reference-based assembly	Trimmed and interleaved reads, and a reference genome
	De novo assembly	Trimmed and interleaved reads, and a seed sequence (coi gene)
NOVOPlasty [Dierckxsens at el 2017]	A seed-extend based assembler	Raw reads, insert size, read length, and a seed sequence (coi gene)
Norgal [Al-Nakeeb at el 2017]	De novo assembly	Raw reads

Outline

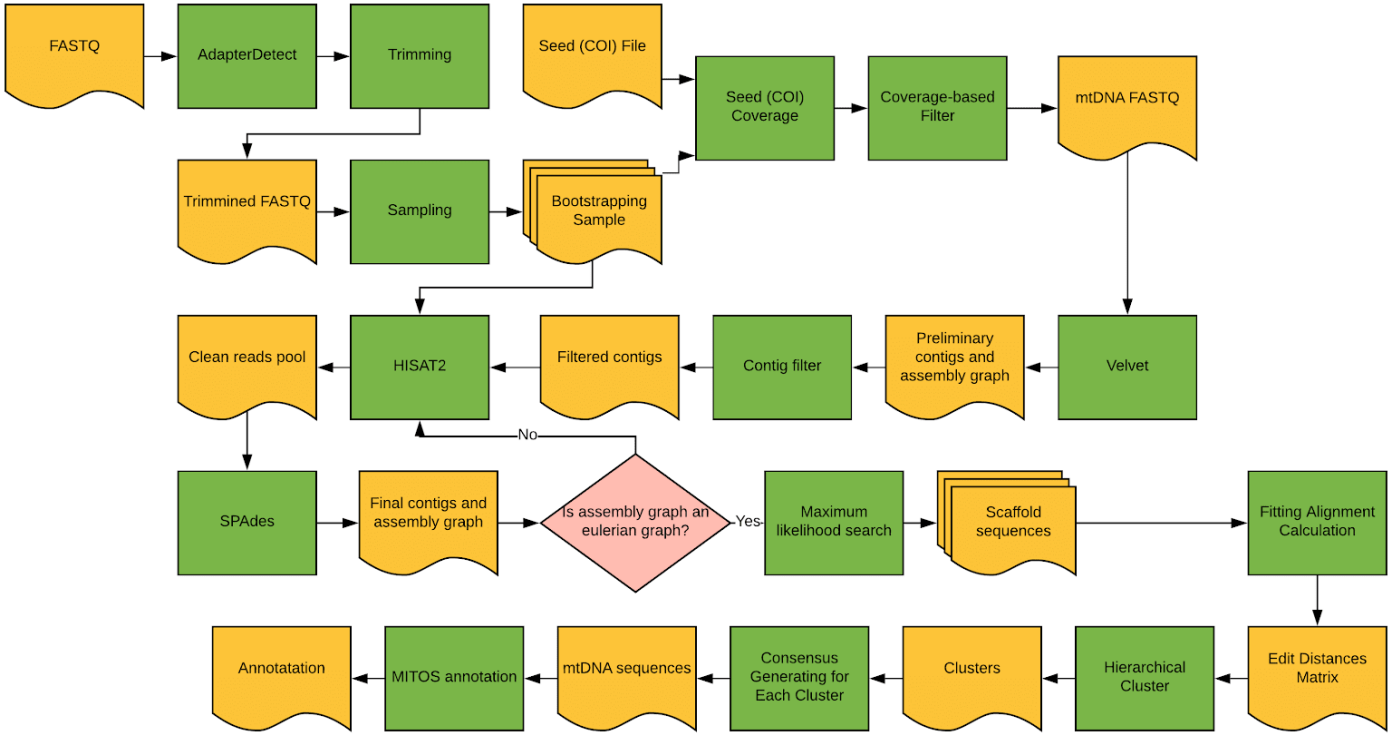
- Background and prior work
- **SMART pipeline**
- Results
- Conclusions and future work

SMART

Statistical **M**itogenome **A**ssembly with **R**epea**T**s

- Input:
  - Paired-end WGS reads
  - Seed sequence (COI gene)
- Output:
  - Complete/circular mitogenome (or largest scaffold)

SMART workflow



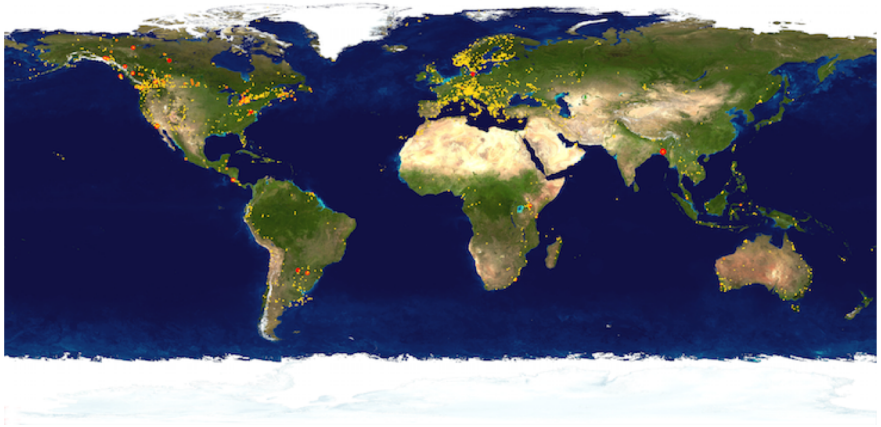
Adapter trimming

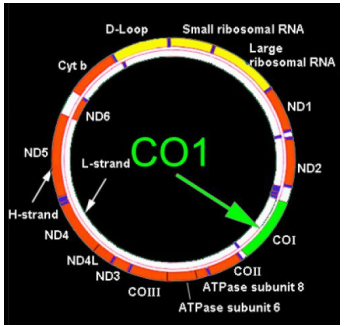
- Automatic detection of adaptors and trimming using Perl/C++ modules from the IRFinder package
  - PE overlap allows very precise (single base resolution) adapter trimming

Middleton, Robert, et al. "IRFinder: assessing the impact of intron retention on mammalian gene expression." Genome biology 18.1 (2017): 51.

Seed (COI) sequences

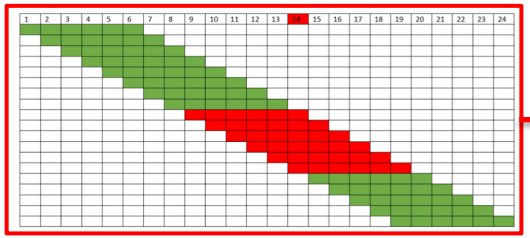
- A ~648bp region of Cytochrome c oxidase subunit 1 (COI) gene has been selected as a “DNA barcode” for taxonomic classification
- Barcode of Life Datasystem (BOLD) has >6M barcodes from 194K animal species, 67K plant species, 21k fungi & other species





http://www.boldsystems.org/

Coverage based filter



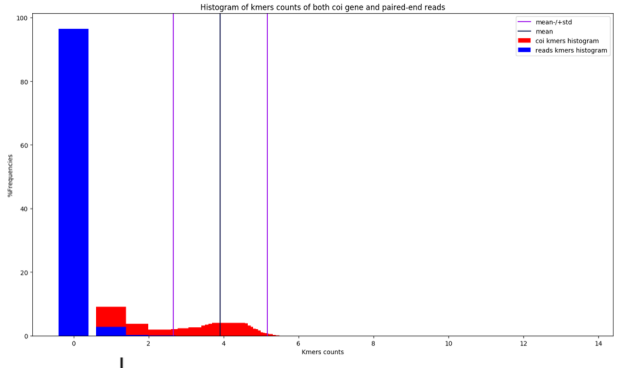
Reads with 1 error OK

Seed (COI)

k-mers

Trimming FASTQ

k-mers & counts



Coverage-based Filter

Hashtable of k-mers with coverage similar to COI

k-mer classifier

Coverage K-mer Classifier:

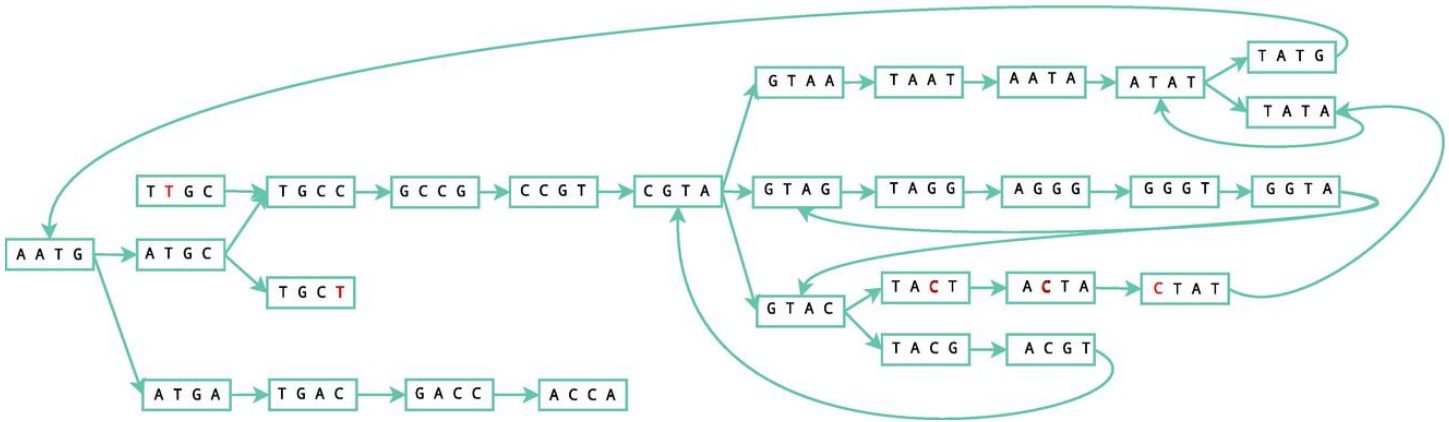
Keep k-mer x if $|\text{coverage}(x) - \mu_{COI}| <= 1\sigma_{COI}$

Keep if ≥l-(2*k)+1 k-mers in hashtable

mtDNA FASTQ

Preliminary assembly

- Reads passing coverage filter assembled using Velvet
  - De Bruijn Graph assembler



https://en.wikipedia.org/wiki/Velvet_assembler

Preliminary contig filtering

- Contigs aligned against eukaryotic mitogenomes using BLAST
  - Keep contigs with significant hits only

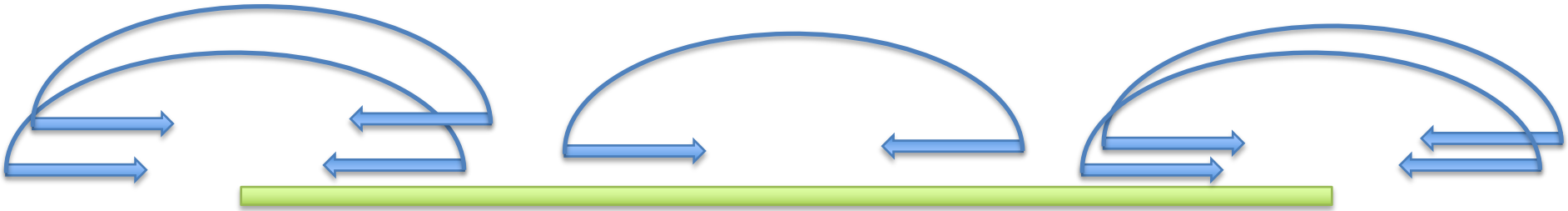
***Eukaryota* mitochondrial genomes - 8376 records**

Alveolata [35]	*Amoebozoa* [9]	*Apusozoa* [0]	*Cryptophyta* [2]
Euglenozoa [0]	*Fornicata* [0]	*Glaucocystophyceae* [4]	*Haptophyceae* [2]
Heterolobosea [5]	*Jakobida* [6]	*Malawimonadidae* [2]	*Opisthokonta* [7957]
Parabasalia [0]	*Rhizaria* [2]	*Rhodophyta* [52]	*Stramenopiles* [83]
Viridiplantae [212]	*unclassified eukaryotes* [5]		

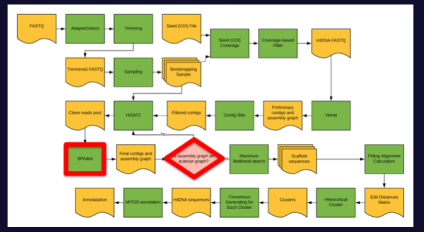
Query label	Target	Percent identity	Alignment length	Number of mismatches	Number of gap	Start position in query	End position in query	Start position in target	End position in target	E-value	Bit score
NODE_1	gi|251831106|ref	99.71	9,753	25	3	1	9,752	9,751	1	0	1.79E+04
NODE_1	gi|251831106|ref	99.69	6,849	21	0	9,753	16,601	16,569	9,721	0	1.25E+04

Read alignment

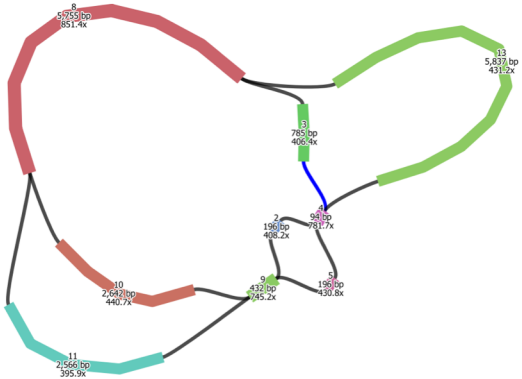
- Using HISAT2
  - Fast and sensitive aligner for NGS reads
- Pulls out additional mitochondrial reads missed by coverage filter



Secondary assembly

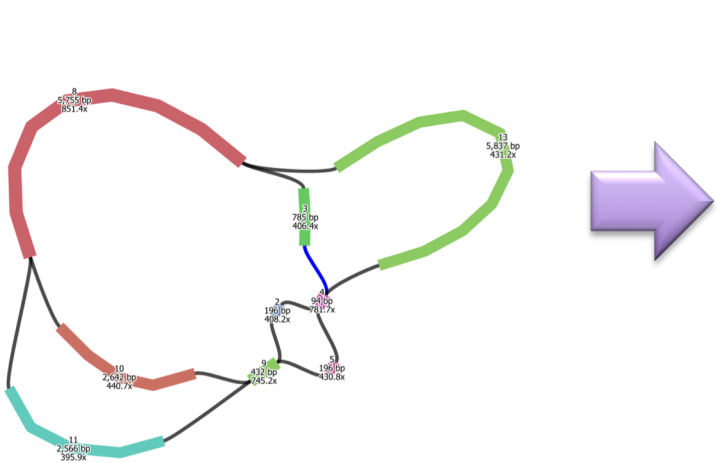


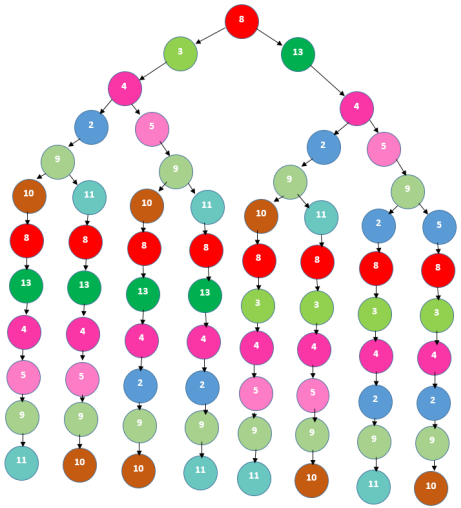
- Using SPAdes
  - Based on multisized de Bruijn graph
  - Robust to non-uniformities in read coverage
- Read alignment and SPAdes assembly repeated
  - Until simplified contig graph is Eulerian, or max iterations reached



Max-likelihood search

- Eulerian paths evaluated using likelihood model implemented in ALE [Clark et al 2013]





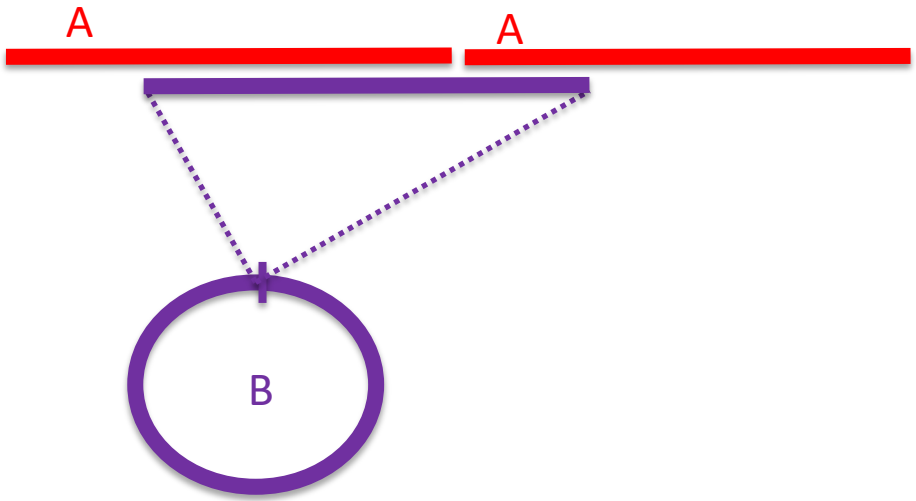
ALE likelihood

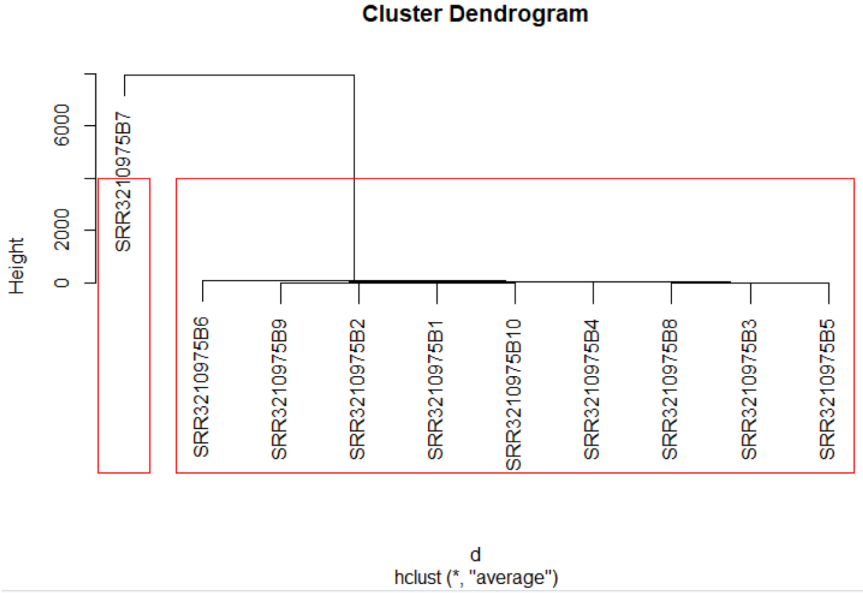
- **Placement scoring:**
  - How well read sequences agree with the assembly
- **Insert scoring:**
  - How well PE insert lengths match those we would expect
- **Depth scoring:**
  - How well depth at each location agrees with depth expected after GC-bias correction
- **K-mer scoring:**
  - How well k-mer counts of each contig match multinomial distribution estimated from entire assembly

https://academic.oup.com/bioinformatics/article/29/4/435/199222

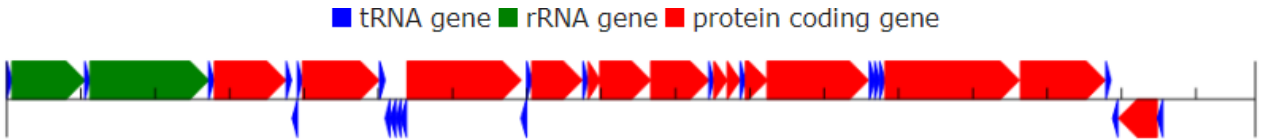
Bootstrapping & clustering

- Process repeated for n=10 bootstrap samples
  - Rotation invariant pairwise distances computed using fitting alignment
  - ML sequences clustered using hierarchical clustering
  - Consensus computed for each cluster



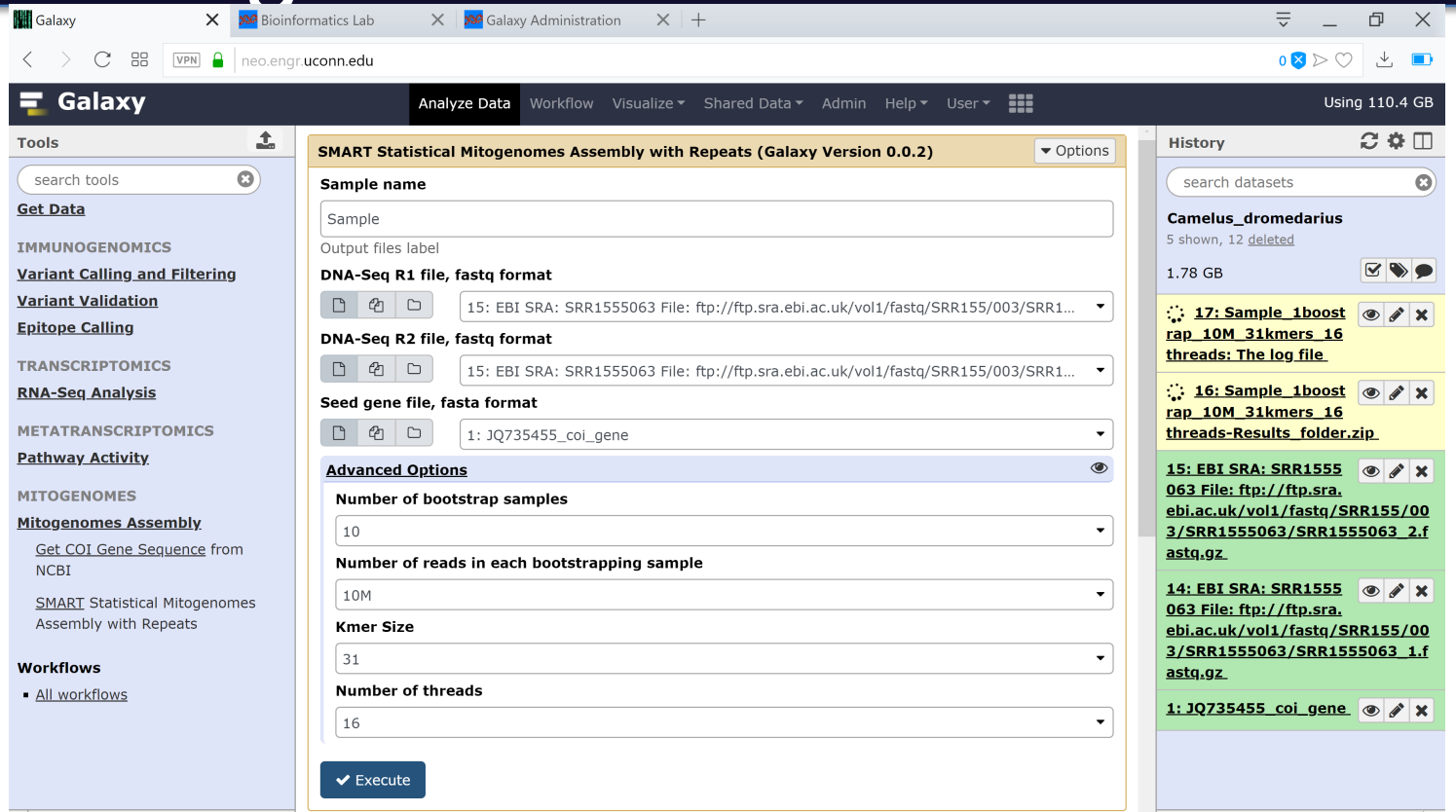


MITOS annotation



Name	Start	Stop	Strand	Length	Structure
trnF(ttc)	1	74	+	74	svg ps
rrnS	74	1053	+	980	svg ps
trnV(gta)	1051	1122	+	72	svg ps
rrnL	1123	2719	+	1597	svg ps
trnL2(tta)	2719	2793	+	75	svg ps
nad1	2798	3754	+	957	
trnI(atc)	3762	3834	+	73	svg ps
trnQ(caa)	3843	3913	-	71	svg ps
trnM(atg)	3913	3981	+	69	svg ps
nad2	3982	5010	+	1029	
trnW(tga)	5021	5091	+	71	svg ps
trnA(gca)	5093	5161	-	69	svg ps
trnN(aac)	5164	5236	-	73	svg ps
trnC(tgc)	5242	5308	-	67	svg ps
trnY(tac)	5309	5378	-	70	svg ps
cox1	5389	6921	+	1533	
trnS2(tca)	6922	6995	-	74	svg ps
trnD(gac)	6999	7067	+	69	svg ps
cox2	7069	7743	+	675	
trnK(aaa)	7754	7823	+	70	svg ps
atp8	7825	7986	+	162	
atp6	7983	8663	+	681	
cox3	8666	9448	+	783	
trnG(gga)	9450	9518	+	69	svg ps
nad3_a	9519	9692	+	174	
nad3_b	9694	9867	+	174	
trnR(cga)	9873	9941	+	69	svg ps
nad4l	9943	10236	+	294	
nad4	10233	11594	+	1362	
trnH(cac)	11611	11680	+	70	svg ps
trnS1(agc)	11681	11747	+	67	svg ps
trnL1(cta)	11748	11817	+	70	svg ps
nad5	11818	13623	+	1806	
cob	13643	14779	+	1137	
trnT(aca)	14790	14859	+	70	svg ps
trnP(cca)	14882	14951	-	70	svg ps
nad6	14962	15480	-	519	
trnE(gaa)	15485	15554	-	70	svg ps

Galaxy interface @ neo.engr.uconn.edu/?toolid=SMART



Outline

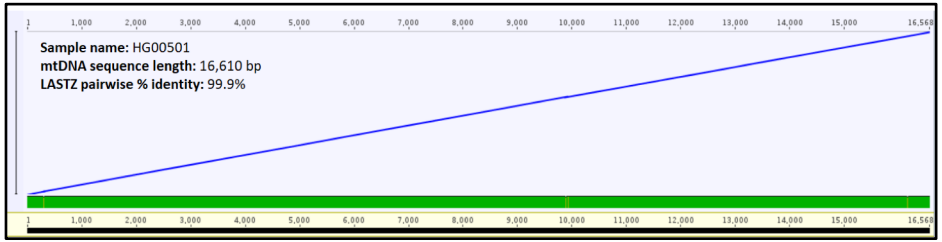
- Background and prior work
- SMART pipeline
- **Results**
- Conclusions and future work

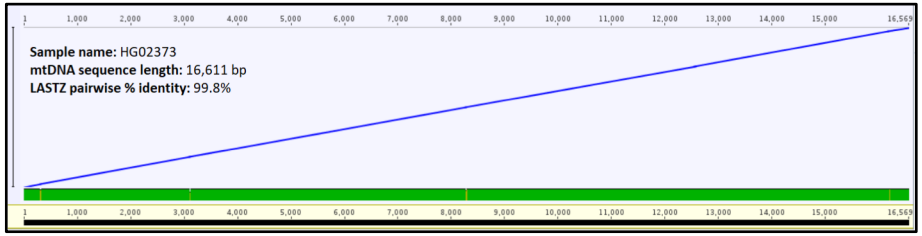
Coverage filter accuracy

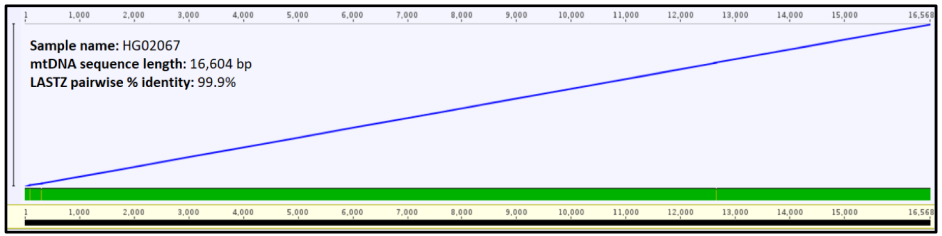
- 2.5M reads
- Ground truth determined by bowtie2 alignment to known reference

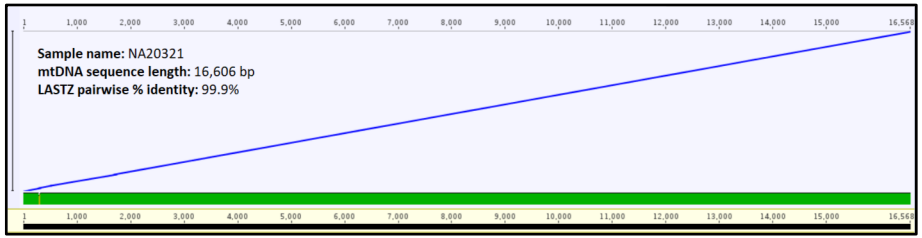
Species	Sample_ID	TPR	PPV	F-Score
Human	HG00501	0.750	0.443	0.557
Human	HG00524	0.454	0.147	0.222
Human	HG00581	0.779	0.516	0.620
Human	HG00635	0.771	0.240	0.366
Chimpanzee	SRR490082	0.715	0.207	0.321
Goat	ERR219544	0.875	0.220	0.352

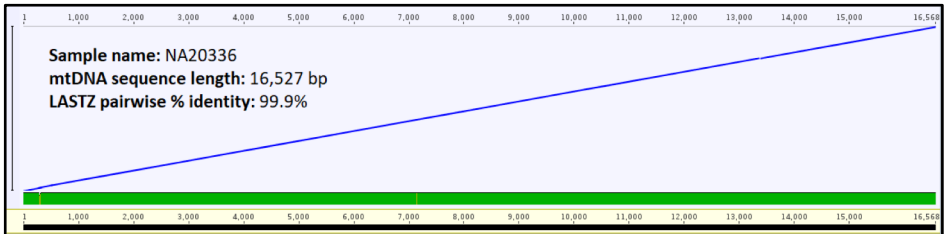
1KGP human datasets











Other datasets

Sample	mtDNA sequence length (bp)	LASTZ pairwise % identity	MUSCLE pairwise % identity	ClustalW pairwise % identity	MAFFT pairwise % identity
Balearica regulorum 	16,742	98.0	98.3	98.3	98.3
Grus japonensis 	16,615	98.4	97.8	97.8	97.8
Xenopus laevis 	17,922	98.0	95.9	96.1	95.7

Other datasets

Sample	mtDNA sequence length (bp)	LASTZ pairwise % identity	MUSCLE pairwise % identity	ClustalW pairwise % identity	MAFFT pairwise % identity
Pan Troglodytes 	16,085	97.5	94.7	94.7	94.7
Mus Musculus 	15,802	99.97	96.9	96.7	96.9
Canis lupus 	16,580	97.1	96.7	96.7	96.7

Other datasets

Sample		mtDNA sequence length (bp)	LASTZ pairwise % identity	MUSCLE pairwise % identity	ClustalW pairwise % identity	MAFFT pairwise % identity
Capra aegagrus hircus		16,098	99.98	96.7	96.7	96.7
Saccharina japonica		37,671	100	99.8	99.8	99.8

Outline

- Background and prior work
- SMART pipeline
- Results
- **Conclusions and future work**

Conclusions

- SMART is an automated pipeline for de novo mitogenome assembly from WGS reads
- Based on statistical framework
  - Probabilistic read classifier based on coverage
  - Likelihood maximization for resolving ambiguities in assembly graph
  - Assembly confidence estimated by bootstrapping
- Produces complete/circular assemblies even in presence of repeats
- Available via galaxy interface at neo.engr.uconn.edu/?toolid=SMART

Ongoing work

- Large-scale pipeline validation
  - 47 frog species from [Zhang et al 2013]
- Comparison with other tools (MITOBim, NOVOPlasty, and Norgal)
- Reconstruction of plant mitochondrial and chloroplast genomes
- Extension to long read sequencing technologies (PacBio, Nanopore)

Thank you for you attention!

Any questions?